

# Rate Matrix Estimation From Site Frequency Data

Conrad J. Burden<sup>1,2</sup> and Yurong Tang<sup>1</sup>

<sup>1</sup> Mathematical Sciences Institute, Australian National University

<sup>2</sup> Research School of Biology, Australian National University

July 18, 2016

## Abstract

A procedure is described for estimating evolutionary rate matrices from observed site frequency data. The procedure assumes (1) that the data are obtained from a constant size population evolving according to a stationary Wright-Fisher model; (2) that the data consist of a multiple alignment of a moderate number of sequenced genomes drawn randomly from the population; and (3) that within the genome a large number of independent, neutral sites evolving with a common mutation rate matrix can be identified. No restrictions are imposed on the scaled rate matrix other than that the off-diagonal elements are positive and  $\ll 1$ , and that the rows sum to zero. In particular the rate matrix is not assumed to be reversible. The key to the method is an approximate stationary solution to the forward Kolmogorov equation for the multi-allele neutral Wright-Fisher model in the limit of low mutation rates.

## 1 Introduction

This paper is a continuation of previous work [1] in which an approximate solution to the forward Kolmogorov equation to the multi-allelic neutral Wright-Fisher model is derived for the biologically relevant case of low mutation rates. Herein we address the problem of estimating a mutation rate matrix from site frequency data. The data is assumed to take the form of a multiple alignment of independent, neutrally evolving genomic sites sequenced from a moderate number of individuals chosen independently from a large effective population.

For an alphabet of size  $K$  alleles the general mutation rate matrix  $Q$  has  $K(K - 1)$  free parameters, which equates to 12 free parameters for the

genomic alphabet  $\{A, C, G, T\}$ . Classical estimates of mutation rates [2, 3], and more recent treatments of the problem (see [4] and references therein) have been concerned primarily with estimating an overall mutation rate, generally denoted by  $\theta$ , whereas the current paper aims to estimate all parameters of the rate matrix  $Q$ . The equivalent estimation problem for  $K = 2$  alleles has been solved by Vogl [5] for neutral sites and Vogl and Bergman [6] when selection is included.

A  $2 \times 2$  rate matrix has a total of 2 free parameters to estimate and is necessarily reversible, which simplifies the problem considerably. The innovation which allows us to deal with the  $K > 2$  cases is an interpretation of the non-reversible part of the rate matrix as a set of fluxes of probability around closed paths in the solution-space simplex of the forward Kolmogorov equation [1]. Section 2 sets out a convention for parameterising the general  $K \times K$  mutation rate matrix  $Q$  which exploits this interpretation. When  $K = 4$ , for instance, we arrive at 3 independent probabilities defining the stationary Markov state, 6 parameters specifying the remaining degrees of freedom in the reversible part of  $Q$ , and 3 probability fluxes specifying the non-reversible part, which sums to the required 12 parameters. Section 3 summarises our previously reported approximate stationary solution to the forward Kolmogorov for the multi-allelic neutral-evolution Wright-Fisher model [1]. Because only low mutation rates are considered the solution can be specified as a set of line densities on the edges and point masses at the corners of the  $(K - 1)$ -dimensional simplex over which the stationary distribution is defined.

The procedure for estimating the parameters of  $Q$  from site frequency data is described in Section 4. Maximum likelihood estimates are obtained assuming the data to consist of counts of allele frequencies observed in a finite sample of individuals assumed to be chosen at random from the population. Interestingly, RoyChoudhury and Wakeley [4] come close to providing the equivalent estimate for the restricted case of a parent-independent rate matrix, but only specify the overall scale  $\theta$  and not the complete rate matrix, which, for their restricted case, has  $K$  parameters and is reversible. Our estimates are tested using synthetic data for  $K = 3$  and  $K = 4$  rate matrices in Section 5. Conclusions are summarised in Section 6.

## 2 Parameterisation of the rate matrix $Q$

Suppose we are given any  $K \times K$  rate matrix  $Q$  whose elements  $Q_{ab}$ , where  $a, b = 1, \dots, K$ , must satisfy

$$Q_{ab} \geq 0, \quad \text{for } a \neq b, \text{ and } \sum_{b=1}^K Q_{ab} = 0. \quad (1)$$

These constraints imply that  $K(K-1)$  parameters are necessary to specify  $Q$ . Inspired by the results of [1] we begin our analysis by constructing a parameterisation consistent with the decomposition of  $Q$  into a reversible part [7, 8] and a flux part, that is,

$$Q = Q^{\text{GTR}} + Q^{\text{flux}}. \quad (2)$$

The flux part represents a set of fluxes of probability around closed paths between subsets of 3 alleles once the Markovian process has settled into its stationary state.

Let us assume that  $Q$  has a unique stationary state  $\pi^T = (\pi_1 \dots \pi_K)$  satisfying

$$\pi_a \geq 0, \quad \sum_{a=1}^K \pi_a = 1, \quad \sum_{a=1}^K \pi_a Q_{ab} = \pi_b. \quad (3)$$

A necessary condition for a unique  $\pi^T$  to exist is that  $Q_{ab} > 0$  for all  $a \neq b$ . One would expect this to include any biologically realistic model. For an evolving population in its stationary state, the rate of mutations from allele- $a$  to allele- $b$  at any genomic site is  $\pi_a Q_{ab}$ .

Define parameters  $C_{ab}$  and  $\Phi_{ab}$  by

$$C_{ab} = \pi_a Q_{ab} + \pi_b Q_{ba}, \quad \Phi_{ab} = \pi_a Q_{ab} - \pi_b Q_{ba}. \quad (4)$$

It is easy to check that

$$Q_{ab} = \frac{1}{2}(C_{ab} + \Phi_{ab})/\pi_a. \quad (5)$$

Hence  $Q$  can be decomposed according to Eq. (2) where

$$Q_{ab}^{\text{GTR}} = \frac{1}{2}C_{ab}/\pi_a, \quad (6)$$

satisfies the time-reversible condition  $\pi_a Q_{ab}^{\text{GTR}} = \pi_b Q_{ba}^{\text{GTR}}$ , and

$$Q_{ab}^{\text{flux}} = \frac{1}{2}\Phi_{ab}/\pi_a. \quad (7)$$

It is clear from Eq. (4) that  $\Phi_{ab}$  is the net flux of probability per unit time from allele- $a$  to allele- $b$ .

Note that there are certain dependencies between the parameters  $\pi_a$ ,  $C_{ab}$  and  $\Phi_{ab}$ . Firstly, the normalisation in Eq. (3) implies that only  $K-1$  components of  $\pi_a$  are independent, i.e.

$$\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i. \quad (8)$$

Secondly,  $C_{ab} = C_{ba}$ , and it follows from the properties of  $Q$  that  $\sum_{b=1}^K C_{ab} = 0$ . Thus  $C_{ab}$  is a symmetric matrix whose diagonal elements are given in terms of its off-diagonal elements via

$$C_{aa} = -\sum_{b \neq a} C_{ab}, \quad a, b = 1, \dots, K. \quad (9)$$

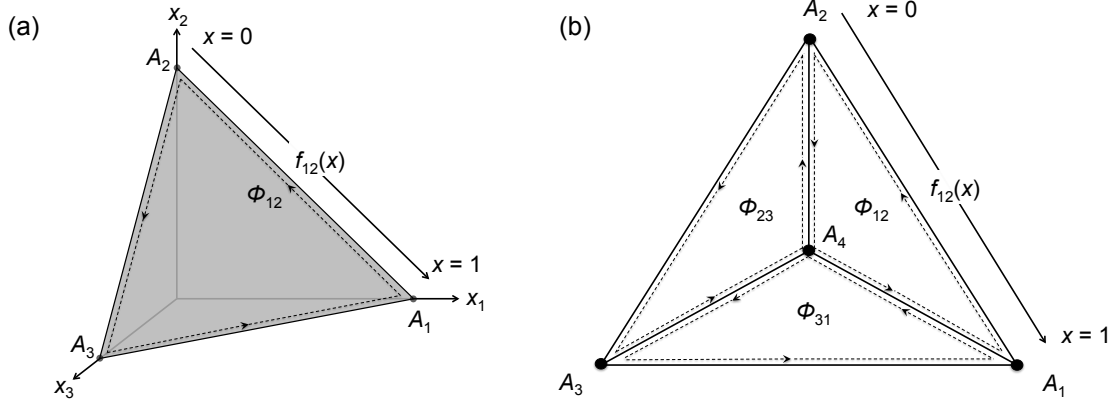


Figure 1: The simplex on which the solution to the forward Kolmogorov equation for the multi-allele Wright-Fisher model is defined for (a)  $K = 3$  alleles and (b)  $K = 4$  alleles. The corners labelled  $A_1$ ,  $A_2$ , etc. indicate the co-ordinates corresponding to non-segregating sites at which the allele specified is prevalent throughout the population. The probability fluxes  $\Phi_{ab}$  and line densities  $f_{ab}(x)$  are explained in the text.

Thirdly,  $\Phi_{ab} = -\Phi_{ba}$ , and it follows from the properties of  $Q$  that  $\sum_{b=1}^K \Phi_{ab} = 0$ . Thus  $\Phi_{ab}$  is an antisymmetric matrix whose rows sum to zero, that is, the final row and column of  $\Phi_{ab}$  are given in terms of the remaining elements via

$$\Phi_{iK} = -\Phi_{Ki} = -\sum_{j \neq i} \Phi_{ij}, \quad i, j = 1, \dots, K-1. \quad (10)$$

Equation (10) is a statement that, in the steady state, the net flux of probability from any allele is zero. For  $K = 3$  alleles there is only one independent flux,  $\Phi_{12}$ , and the elements of  $Q$  are

$$Q = \frac{1}{2} \begin{pmatrix} \frac{-C_{12} - C_{13}}{\pi_1} & \frac{C_{12} + \Phi_{12}}{\pi_1} & \frac{C_{13} - \Phi_{12}}{\pi_1} \\ \frac{C_{12} - \Phi_{12}}{\pi_2} & \frac{-C_{12} - C_{23}}{\pi_2} & \frac{C_{23} + \Phi_{12}}{\pi_2} \\ \frac{C_{13} + \Phi_{12}}{1 - \pi_1 - \pi_2} & \frac{C_{23} - \Phi_{12}}{1 - \pi_1 - \pi_2} & \frac{-C_{13} - C_{23}}{1 - \pi_1 - \pi_2} \end{pmatrix}. \quad (11)$$

For  $K = 4$  alleles there are three independent fluxes  $\Phi_{12}$ ,  $\Phi_{23}$  and  $\Phi_{31}$  as illustrated in Fig. 1.

To summarise, the general rate matrix  $Q$  can be parameterised via

Eqs. (2), (6) and (7) using the following minimal set of parameters:

$$\begin{aligned} \pi_i, & \quad i = 1, \dots, K-1 : & K-1 \text{ parameters;} \\ C_{ab} = C_{ba}, & \quad 1 \leq a < b \leq K : & \frac{1}{2}K(K-1) \text{ parameters;} \\ \Phi_{ij} = -\Phi_{ji}, & \quad 1 \leq i < j \leq K-1 : & \frac{1}{2}(K-1)(K-2) \text{ parameters,} \end{aligned} \quad (12)$$

with the remaining, unspecified parameters given by Eqs. (8), (9) and (10). The total number of independent parameters listed in Eq. (12) is  $K(K-1)$ , as required. The requirement that the off-diagonal elements of  $Q$  be positive implies the further constraints on the parameter space that

$$\pi_a \geq 0, \quad C_{ab} \geq 0, \quad |\Phi_{ab}| \leq C_{ab}, \quad 1 \leq a < b \leq K. \quad (13)$$

The remainder of this paper is concerned with estimating the  $K(K-1)$  parameters of a genomic evolutionary rate matrix from site frequency data assuming a population whose genome includes a large number of independent sites that have evolved to stationarity according to a neutral evolution Wright-Fisher model.

### 3 Approximate solution to the neutral multi-allele Wright-Fisher model

We consider the neutral evolution Wright-Fisher model for  $K$  alleles, labelled  $A_1 \dots A_K$  (see, for example, Section 4.1 of ref. [9]). Given a haploid population of size  $N$  (or monoecious diploid population of size  $N/2$ ), let the number of individuals of type  $A_a$  at time step  $\tau$  be  $Z_a(\tau)$  for discrete times  $\tau = 0, 1, 2, \dots$ . Also, let  $u_{ab}$  be the probability of an individual making a transition from  $A_a$  to  $A_b$  in a single time step, where  $u_{ab} \geq 0$  and  $\sum_{b=1}^K u_{ab} = 1$ . Writing  $\mathbf{Z}(\tau) = (Z_1(\tau), \dots, Z_K(\tau))$ , the multi-allele neutral Wright-Fisher model is defined by the transition matrix from an allele frequency  $\mathbf{i} = (i_1, \dots, i_K)$  to an allele frequency  $\mathbf{j} = (j_1, \dots, j_K)$  in the population given by

$$\text{Prob}(\mathbf{Z}(\tau+1) = \mathbf{j} | \mathbf{Z}(\tau) = \mathbf{i}) = \frac{N!}{\prod_{a=1}^K j_a!} \prod_{a=1}^K \psi(\mathbf{i}, a)^{j_a}, \quad (14)$$

where  $\sum_{a=1}^K i_a = \sum_{a=1}^K j_a = N$ , and

$$\psi(\mathbf{i}, a) = \frac{i_a}{N} \left( 1 - \sum_{b \neq a} u_{ab} \right) + \sum_{b \neq a} \frac{i_b}{N} u_{ba} = \sum_{b=1}^K \frac{i_b}{N} u_{ba}. \quad (15)$$

The usual diffusion limit is obtained by defining random variables  $X_a(t) = Z_a(\tau)/N$  equal to the relative proportion of type- $A_a$  alleles within the population at continuous time  $t = \tau/N$ . The limit  $N \rightarrow \infty$  and  $u_{ab} \rightarrow 0$  for

$a \neq b$  is taken in such a way that the  $K \times K$  instantaneous rate matrix  $Q$ , whose elements are defined by

$$Q_{ab} = N(u_{ab} - \delta_{ab}), \quad (16)$$

remains finite. This limit leads to a forward Kolmogorov equation for the density function  $f_{\mathbf{X}}(x_1, \dots, x_{K-1}; t)$  of the vector of continuous random variables  $X_1(t), \dots, X_{K-1}(t)$ . The function  $f_{\mathbf{X}}$  is defined over the simplex (see Fig. 1)

$$\mathcal{S} = \left\{ (x_1, \dots, x_K) : x_1, \dots, x_K \geq 0, \sum_{a=1}^K x_a = 1 \right\}. \quad (17)$$

Further details of the equation are summarised in [1].

Solution of the forward Kolmogorov equation for an arbitrary rate matrix  $Q$  and  $K \geq 3$  alleles, even for the stationary distribution when  $\partial f_{\mathbf{X}}/\partial t$  is set to zero, remains an unsolved problem. However, in [1] we derive an approximate stationary distribution in the biologically realistic limit of slow but otherwise arbitrary mutation rates, that is for  $0 \leq Q_{ab} \ll 1$  for  $a \neq b$ . Our analysis is based on the result that, in the limit  $Q_{ab} \rightarrow 0$ , the stationary probability distribution is concentrated close to the edges of the simplex  $\mathcal{S}$ , and can therefore be represented accurately as a set of line densities defined along those edges. Suppose we label the corners of  $\mathcal{S}$  by the allele prevalent within the population at that corner, so the corner  $A_1$  corresponds to the co-ordinate  $(x_1, \dots, x_K) = (1, 0, \dots, 0)$ , and so on, as illustrated in Fig. 1. On the edge joining corner  $A_a$  to corner  $A_b$  define a line density  $f_{ab}(x)$  for each pair of indices  $a$  and  $b$ . We will adopt the convention that the argument  $x$  is the relative proportion of type- $a$  alleles, and  $1 - x$  is the relative proportion of type- $b$  alleles. The relative proportion of the remaining  $K - 2$  alleles along this edge is zero.

The line densities are given in terms of the parameterisation introduced in Section 2 as (see Eq. (53) of [1])

$$f_{ab}(x) = C_{ab} \left( \frac{1}{x} + \frac{1}{1-x} \right) - \Phi_{ab} \left( \frac{1}{x} - \frac{1}{1-x} \right). \quad (18)$$

Note that  $f_{ab}(x) = f_{ba}(1 - x)$ . A necessary condition for the line density to be an accurate representation of the exact solution is that

$$Q_{ab} \times |\log(x) + \log(1 - x)| \ll 1. \quad (19)$$

Thus the approximation loses accuracy as  $x$  approaches 0 or 1, that is, in the vicinity of the corners of  $\mathcal{S}$ . However, for a population of size  $N$ , the value of the stationary distribution at the corners of the simplex corresponding to the discrete problem defined by Eq. (14) is, to a good approximation, (see Eq. (56) of [1])

$$P(A_a) = \text{Prob}(Z_a = N, Z_b = 0 \text{ for } b \neq a) = \pi_a - \sum_{b \neq a} C_{ab} \log N. \quad (20)$$

As a rule of thumb, we have observed in numerical simulations that Eqs. (18) and (20) provide a very good approximation to the stationary state of the discrete model provided the off-diagonal elements of  $Q$  are less than  $10^{-2}$ . The approximate solution is normalised in the sense that

$$\lim_{N \rightarrow \infty} \left( \sum_{1 \leq a < b \leq K} \int_{1/N}^{1-1/N} f_{ab}(x) dx + \sum_{a=1}^K P(A_a) \right) = 1. \quad (21)$$

## 4 Parameter Estimation

Assume we have a data set in the form of a site frequency spectrum (SFS) obtained by sampling  $L$  independent neutrally evolving sites within the genomes of  $M$  individuals from a population of size  $N \gg M$ . Typically  $L$  might be at least  $10^3$ ,  $M$  in the range 10 to 100, and  $N$  is ideally essentially infinite in the sense that the diffusion limit forward-Kolmogorov equation is appropriate.  $L$ ,  $M$  and  $N$  are known fixed parameters. At each genomic site  $l$ , define a vector of non-negative integer valued random variables

$$\mathbf{Y}^{(l)} = Y_1^{(l)}, \dots, Y_K^{(l)}, \quad l = 1, \dots, L, \quad (22)$$

where  $Y_a^{(l)}$  equal to the number of times allele type- $a$  occurs within the sampled individuals. Clearly  $\sum_{a=1}^K Y_a^{(l)} = M$ , so the data at any given genomic site can be specified as a point in a  $(K-1)$ -dimensional simplex lattice.

Given an observed data set  $(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(L)})$ , the log-likelihood is

$$\mathcal{L}(\pi, C, \Phi | \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(L)}) = \sum_{l=1}^L \log \text{Prob}(\mathbf{Y}^{(l)} = \mathbf{y}^{(l)} | \pi, C, \Phi), \quad (23)$$

where the triplet  $(\pi, C, \Phi)$  represents the  $K(K-1)$  parameters of Eq. (12). The probabilities occurring in this sum are calculated under the assumption that the data is sampled randomly from a population with genomic sites distributed according to the approximate stationary solution of Section 3. Below we show that these probabilities are given by

$$\text{Prob}(\mathbf{Y}^{(l)} = \mathbf{y} | \pi, C, \Phi) = \begin{cases} \pi_a - H_M \sum_{b \neq a} C_{ab}, & \text{if } y_a = M \text{ and all other} \\ & \text{components of } \mathbf{y} \text{ are zero,} \\ C_{ab} \left( \frac{1}{y} + \frac{1}{M-y} \right) - \Phi_{ab} \left( \frac{1}{y} - \frac{1}{M-y} \right) & \text{if } y_a = y, y_b = M - y \text{ and all} \\ & \text{other components of } \mathbf{y} \text{ are zero,} \\ 0 & \text{otherwise,} \end{cases} \quad (24)$$

where

$$H_M = \sum_{y=1}^{M-1} \frac{1}{y}. \quad (25)$$

This distribution generalises the corresponding  $K = 2$  distribution found previously by Vogl, namely Eq. (13) of [5], for which Vogl and Bergman [6] propose the name “generalised RoyChoudhury-Wakeley distribution”. There is no  $\Phi_{ab}$  contribution for  $K = 2$  as any  $2 \times 2$  rate matrix is automatically reversible. Furthermore RoyChoudhury and Wakeley’s distribution, namely Eq. (10) of [4], is actually the restriction of the second part of Eq. (24) to the case of a  $K$ -allele parent-independent rate matrix,  $Q_{ab} = \frac{1}{2}\theta\pi_b$ , for arbitrary  $K \geq 2$  and a canonical scaling parameter  $\theta \ll 1$ . In terms of our parameterisation Eq. (12) this corresponds to setting  $C_{ab} = \theta\pi_a\pi_b$  and  $\Phi_{ij} = 0$ . The parent-independent rate matrix is also reversible, and includes the most general  $2 \times 2$  rate matrix when  $K = 2$ . We therefore propose that the name generalised RoyChoudhury-Wakeley-Vogl-Bergman distribution could be appropriately applied to Eq. (24). A proof of Eq. (24) follows.

*Proof.* Consider the three cases in turn.

**Case I:** Sites for which exactly 1 component of  $\mathbf{y}^{(l)}$  is non-zero. Suppose site- $l$  is non-segregating with allele  $A_a$  occurring in all individuals within the sample. In this case one can reduce the continuum diffusion limit to an effective 2-allele model in which all alleles except  $A_a$  are combined into a single allele,  $A_{\bar{a}}$ . As described in Eqs. (A.4) and (A.5) of Appendix A of [1], the effective 2-allele model has a rate matrix

$$\tilde{Q} = \begin{pmatrix} 1 - \tilde{Q}_{a\bar{a}} & \tilde{Q}_{a\bar{a}} \\ \tilde{Q}_{\bar{a}a} & 1 - \tilde{Q}_{\bar{a}a} \end{pmatrix}, \quad (26)$$

with,

$$\begin{aligned} \tilde{Q}_{a\bar{a}} &= \frac{\sum_{b \neq a} \pi_a Q_{ab}}{\pi_a} = \frac{\sum_{b \neq a} C_{ab}}{2\pi_a}, \\ \tilde{Q}_{\bar{a}a} &= \frac{\sum_{b \neq a} \pi_b Q_{ba}}{1 - \pi_a} = \frac{\sum_{b \neq a} C_{ab}}{2(1 - \pi_a)} \end{aligned} \quad (27)$$

where we have used Eq. (5) and the fact that  $\Phi_{ab}$  is an anti-symmetric matrix whose rows sum to zero. In Section 4 of Ref. [5] Vogl provides an analysis of the 2-allele model in the small mutation rate limit. Vogl parameterises the  $2 \times 2$  rate matrix in terms of two parameters,  $\vartheta$  and  $\alpha$ , which are related to our parameters by  $\tilde{Q}_{a\bar{a}} = \vartheta/(2\alpha)$ ,  $\tilde{Q}_{\bar{a}a} = \vartheta/[2(1 - \alpha)]$ , or equivalently,

$$\vartheta = \sum_{b \neq a} C_{ab}, \quad \alpha = \pi_a. \quad (28)$$



From Eq. (29) of [5], using the above identification we can immediately read off the required probability

$$\text{Prob} \left\{ \mathbf{Y}^{(l)} = (0, \dots, Y_a^{(l)} = M, \dots, 0) \right\} = \pi_a - H_M \sum_{b \neq a} C_{ab}. \quad (29)$$

**Case II:** Sites for which exactly 2 components of  $\mathbf{y}^{(l)}$  are non-zero. Suppose site  $l$  is biallelic with alleles  $A_a$  and  $A_b$  occurring  $y_a^{(l)} = y$  times and  $y_b^{(l)} = M - y$  times respectively within the sample. Then, from Eq. (18),

$$\begin{aligned} & \text{Prob} \left\{ \mathbf{Y}^{(l)} = (0, \dots, y, \dots, M - y, \dots, 0) \right\} \\ &= \int_{1/N}^{1-1/N} f_{ab}(x) \binom{M}{y} x^y (1-x)^{M-y} dx \\ &= \binom{M}{y} \left[ (C_{ab} - \Phi_{ab}) \int_{1/N}^{1-1/N} x^{y-1} (1-x)^{M-y} dx \right. \\ & \quad \left. + (C_{ab} + \Phi_{ab}) \int_{1/N}^{1-1/N} x^y (1-x)^{M-y-1} dx \right] \\ &= \binom{M}{y} [(C_{ab} - \Phi_{ab}) B(y, M - y + 1) \\ & \quad + (C_{ab} + \Phi_{ab}) B(y + 1, M - y)] + O\left(\frac{1}{N}\right) \end{aligned} \quad (30)$$

where  $B(m, n) = \Gamma(m)\Gamma(n)/\Gamma(m+n)$  is the beta function. For positive integer arguments  $\Gamma(n) = (n-1)!$ , which reduces the last line to

$$\begin{aligned} & \text{Prob} \left\{ \mathbf{Y}^{(l)} = (0, \dots, y, \dots, M - y, \dots, 0) \right\} \\ & \approx C_{ab} \left( \frac{1}{y} + \frac{1}{M - y} \right) - \Phi_{ab} \left( \frac{1}{y} - \frac{1}{M - y} \right), \end{aligned} \quad (31)$$

up to order  $1/N$ .

**Case III:** Sites for which 3 or more components of  $\mathbf{y}^{(l)}$  are non-zero. The assumption that the solution to the forward Kolmogorov equation can be represented by a set of line densities on the edges plus point masses at the vertices implies that, in the entire the population, no more than two alleles can be represented at any given genomic site. In other words, the assumed model entails that this case will occur with probability zero.  $\square$

Regarding Case III, since 3-allelic and 4-allelic sites are rare in genomes [10, 11], we argue that removing such sites from the data will not do serious damage to a maximum likelihood estimator of the parameters. Alternatively, one could reassign such data points to the nearest point on an edge of the simplex lattice.

Now define the following variables:

$$\begin{aligned}
L_a &= \sum_{i=1}^L I \left( Y_a^{(i)} = M, Y_b^{(i)} = 0 \text{ for } b \neq a \right), \quad a = 1, \dots, K, \\
L_{ab}(y) &= \sum_{i=1}^L I \left( Y_a^{(i)} = y, Y_b^{(i)} = M - y, Y_c^{(i)} = 0 \text{ for } c \neq a, b \right), \\
&\quad 1 \leq a < b \leq K; y = 1, \dots, M - 1, \\
L_{ab} &= \sum_{y=1}^{M-1} L_{ab}(y),
\end{aligned} \tag{32}$$

where  $I(\cdot)$  is the indicator random variable for the event specified. That is,  $L_a$  is a count of the number of non-segregating sites of allele type  $A_a$ ,  $L_{ab}(y)$  is a count of the number of biallelic polymorphisms with  $y$  occurrences of allele  $A_a$  and  $M - y$  occurrences of allele  $A_b$ , and  $L_{ab}$  is the total number of biallelic polymorphisms of type  $A_a$ - $A_b$ . We will assume the data is such that all sites observed are either non-segregating or biallelic, i.e.,  $\sum_a L_a + \sum_{a < b} L_{ab} = L$ . Then

$$\begin{aligned}
\text{Prob}(L_a = l_a, L_{ab}(y) = l_{ab}(y) \mid \pi, c, \Phi) = \\
\frac{L!}{(\prod_{a=1}^K l_a!)(\prod_{a < b} \prod_{y=1}^{M-1} l_{ab}(y)!)} \left[ \prod_{a=1}^K \left( \pi_a - H_M \sum_{b \neq a} C_{ab} \right)^{l_a} \right] \times \\
\left[ \prod_{a < b} \prod_{y=1}^{M-1} \left( C_{ab} \left\{ \frac{1}{y} + \frac{1}{M-y} \right\} - \Phi_{ab} \left\{ \frac{1}{y} - \frac{1}{M-y} \right\} \right)^{l_{ab}(y)} \right],
\end{aligned} \tag{33}$$

and

$$\begin{aligned}
\text{Prob}(L_a = l_a, L_{ab} = l_{ab} \mid \pi, c, \Phi) = \\
\frac{L!}{(\prod_{a=1}^K l_a!)(\prod_{a < b} l_{ab}!)} \left[ \prod_{a=1}^K \left( \pi_a - H_M \sum_{b \neq a} C_{ab} \right)^{l_a} \right] \left[ \prod_{a < b} (2H_M C_{ab})^{l_{ab}} \right].
\end{aligned} \tag{34}$$

Since Eq. (34) does not depend on  $\Phi$ , the observations  $L_a$  and  $L_{ab}$  are sufficient statistics for estimating  $\pi_i$  and  $C_{ab}$  [3, 4]. Moreover, for any index  $a = 1, \dots, K$ , one can again define an effective 2-allele model by partitioning the set of alleles into  $A_a$  and an effective allele  $A_{\bar{a}}$  consisting of the remaining  $K - 1$  alleles, and then use Vogl's unbiased maximum-likelihood estimator

for  $\hat{\alpha}$  (see Eq. (37) of [5]) together with Eq. (28) to obtain the estimators

$$\hat{\pi}_a = \frac{1}{L} \left( L_a + \frac{1}{2} \sum_{b \neq a} L_{ab} \right). \quad (35)$$

Similarly, one can consider a broader set of partitionings of alleles into two exhaustive disjoint subsets and the corresponding effective 2-allele models, together with Vogl's unbiased maximum-likelihood estimator for  $\hat{v}$  (see Eq. (36) of [5]) to obtain the estimators

$$\hat{C}_{ab} = \frac{L_{ab}}{2LH_M}. \quad (36)$$

It is a straightforward exercise to confirm using Eqs.(24) and (32) that these are unbiased estimators.

It remains to estimate the flux parameters  $\Phi_{ij}$ . In the following we carry out a numerical maximisation of the log-likelihood formed from Eq. (33) by plugging in the above estimates of  $\hat{C}_{ab}$ . Up to an additive constant this gives

$$\begin{aligned} & \mathcal{L}(\Phi_{ij} | l_a, l_{ab}(1), \dots, l_{ab}(M-1)) \\ &= \sum_{a < b} \sum_{y=1}^{M-1} l_{ab}(y) \log \left( \hat{C}_{ab} \left\{ \frac{1}{y} + \frac{1}{M-y} \right\} - \Phi_{ab} \left\{ \frac{1}{y} - \frac{1}{M-y} \right\} \right) \end{aligned} \quad (37)$$

The right hand side of this formula is a function of  $\frac{1}{2}(K-1)(K-2)$  parameters  $\Phi_{ij}$  for  $1 \leq i < j \leq K-1$ , with Eq. (10) used to interpret the terms in the sum for which  $b = K$ .

## 5 Results

We have constructed a number of synthetic datasets to test the efficacy of the above theory. Since the exact solution of the Forward Kolmogorov equation is unknown, starting from an assumed scaled rate matrix  $Q$  we first generate numerically a stationary site frequency spectrum from the neutral Wright-Fisher model for as large a population  $N$  as is practicable. For this step the full transition matrix, Eq. (14), of size  $\binom{N+K-1}{K-1} \times \binom{N+K-1}{K-1}$  is used. Each dataset is then created corresponding to a multiple alignment at a large number of  $L$  independent genomic sites of the genomes of  $M$  individuals sampled randomly from the population. Here we have aimed to satisfy the ideal limit  $M \ll N$  within the constraints of the simulation. Each genomic site is assumed to be the result of the Markov process of the full transition matrix evolving to its stationary state. Details of the sampling process are described in detail below. Finally the parameters of  $Q$  for each of 1000 such independently generated datasets are estimated using the theory of Section 4, and the results presented as histograms.

## 5.1 Synthetic Data: $K = 3$ Alleles

For  $K = 3$  alleles, and for each rate matrix tested, a numerical stationary solution to the neutral Wright-Fisher model with matrix Eq. (14) was first created for a population size  $N = 100$ . Three rate matrices were considered. Each matrix had the same reversible part,  $Q^{\text{GTR}}$ , corresponding to the parameters (see Eq. (11))

$$(\pi_1, \pi_2) = (0.5, 0.3), \quad (C_{12}, C_{23}, C_{13}) = (0.0003, 0.0006, 0.0004). \quad (38)$$

For the single flux parameter which determines  $Q^{\text{flux}}$ , namely  $\Phi \equiv \Phi_{12}$ , three cases were considered:

$$\Phi = 0, \quad \Phi = 0.0001, \quad \text{and} \quad \Phi = 0.0002. \quad (39)$$

For each value of  $\Phi$  total of 1000 synthetic datasets were constructed, each assuming a sample of  $M = 10$  individuals sequenced at  $L = 10^5$  independent genomic sites. At each site  $l \in \{1, \dots, L\}$  and within each dataset the stationary distribution was sampled to establish the relative frequencies of the 3 alleles in the population at that site. These relative frequencies were used as parameters of a multinomial distribution from which we sampled the observed counts  $y_a^{(l)} \in \{0, \dots, M\}$  of the number of times allele  $A_a$  was observed at site  $l$ . Any genomic site displaying more than 2 alleles in the sample was discarded, though generally this amounted to no more than 1 or 2 tri-allelic SNPs observed per dataset. Maximum likelihood estimates  $\hat{\pi}_a$ ,  $\hat{C}_{ab}$  and  $\hat{\Phi}$  of the parameters defining  $Q$  were obtained for each dataset using the theory developed in Section 4 (see Eqs. (35) to (37)).

Histograms of the estimates  $\hat{\pi}_a$  and  $\hat{C}_{ab}$  are shown in Fig. 2. The estimates  $\hat{\pi}_a$  are centred about their true values, while the estimates  $\hat{C}_{ab}$  tend to be slightly low, perhaps because of the approximate nature of the solution to the forward Kolmogorov equation. The distributions of these parameters are independent of  $\Phi$ , as expected.

Histograms of the estimates  $\hat{\Phi}$  are shown in Fig. 3. These estimates are well centred about their true values. Under the null hypothesis that  $\Phi = 0$  the likelihood ratio test statistic

$$-2 \left[ \mathcal{L}(0 | l_a, l_{ab}(1), \dots, l_{ab}(M-1)) - \mathcal{L}(\hat{\Phi}_{ij} | l_a, l_{ab}(1), \dots, l_{ab}(M-1)) \right], \quad (40)$$

where  $\mathcal{L}(\cdot)$  is given by Eq. (37), should have a chi-squared distribution with 1 degree of freedom. For each true value of  $\Phi$  and each dataset a one-sided p-value was calculated from this statistic using the upper tail of the chi-squared distribution. Shown in the right-hand panel of Fig. 3 for each  $\Phi$ -value are ordered p-values from the 1000 datasets, plotted against quantiles of a uniform distribution. For the  $\Phi = 0$  datasets the p-values have a uniform distribution as required, while the p-values for the remaining two values of  $\Phi$  are suitably small.

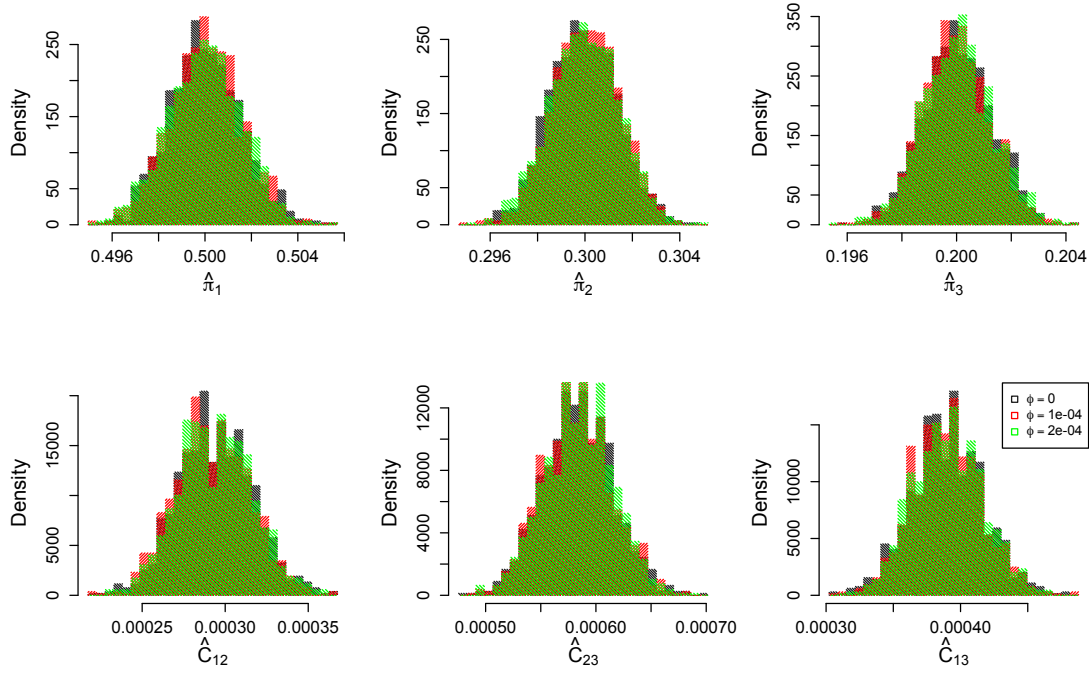


Figure 2: Histograms of estimates  $\hat{\pi}_a$  and  $\hat{C}_{ab}$  from of the parameters defining the reversible part  $Q^{\text{GTR}}$  of the rate matrix for  $K = 3$  alleles. True values of the parameters are  $(\pi_1, \pi_2, \pi_3) = (0.5, 0.3, 0.2)$  and  $(C_{12}, C_{23}, C_{13}) = (0.0003, 0.0006, 0.0004)$ . 1000 independent datasets were generated for each of 3 values of the flux parameter,  $\Phi = 0, 0.0001$  and  $0.0002$ .

## 5.2 Synthetic Data: $K = 4$ Alleles

A numerical stationary solution to the neutral Wright-Fisher model with transition matrix Eq. (14) was created for a population size  $N = 30$  for each of the following 4 rate matrices, defined by the parameters in Table 1:

**GTR.** The  $4 \times 4$  general time reversible matrix has 9 independent parameters, which can be specified as the  $\pi_i$  and  $C_{ab}$  defined in Eq. (12). The first 9 parameters in the first column of Table 1 are chosen to reproduce very approximately the time-reversible matrix in Table 5 of [7] estimated from the rat-mouse phylogeny. The resulting scaled GTR matrix is

$$Q^{\text{GTR}} = \begin{pmatrix} -5.375 & 1.875 & 2.000 & 1.5000 \\ 2.500 & -17.500 & 0.333 & 14.667 \\ 16.000 & 2.000 & -21.000 & 3.000 \\ 2.400 & 17.600 & 0.600 & -20.600 \end{pmatrix} \times 10^{-4}. \quad (41)$$

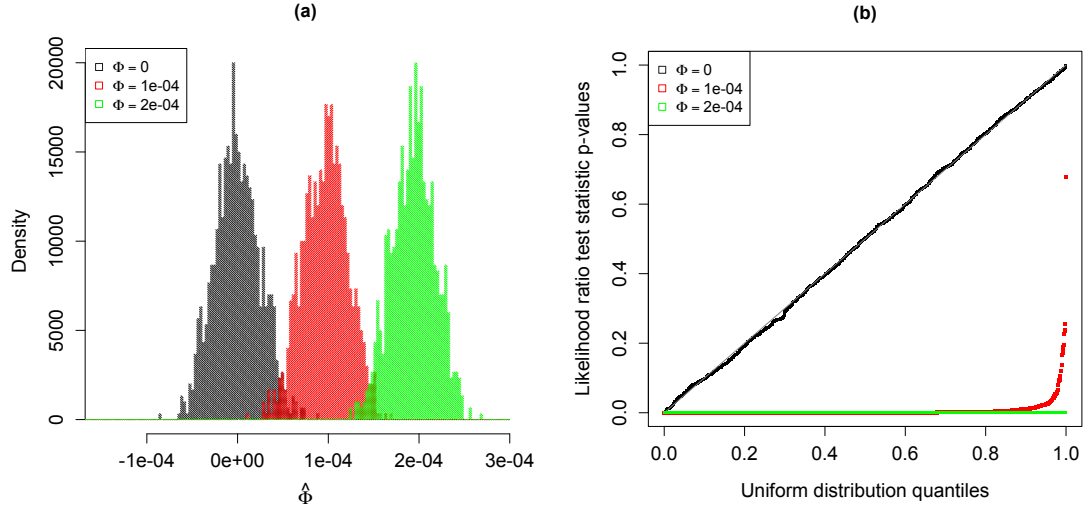


Figure 3: (a) Histograms of estimates  $\hat{\Phi}$  from of the flux parameter contributing to the non-reversible part  $Q^{\text{flux}}$  of the rate matrix for  $K = 3$  alleles. True values of the parameters are as in the caption to Fig. 2. (b) Ordered 1-sided p-values calculated from the likelihood ratio test statistic Eq. (40) plotted against uniform quantiles. For each true value of  $\Phi$  p-values are calculated from 1000 synthetic datasets.

Note that Lavane et al's rate matrix is a per-generation rate and differs from our scaled matrix by a factor of  $10^5$ , which, by Eq. (16), corresponds to assuming the effective population size of Lanave et al's data set to be  $10^5$ .

**GRM.** The most general  $4 \times 4$  rate matrix has 12 parameters, which can be specified as the parameters  $\pi_i$ ,  $C_{ab}$  and  $\Phi_{ij}$  defined in Eq. (12). The general rate matrix  $Q^{\text{GRM}}$  in our simulations uses the same reversible part as  $Q^{\text{GTR}}$ , plus a non-reversible part using the  $\Phi_{ij}$  specified in the first column of Table 1, which are chosen arbitrarily subject to the constraint that they should lie within the allowable range specified by Eq. (13). This gives

$$Q^{\text{GRM}} = \begin{pmatrix} -5.375 & 3.125 & 2.125 & 0.1250 \\ 0.833 & -17.500 & 0.583 & 16.083 \\ 15.000 & 0.500 & -21.000 & 5.500 \\ 4.600 & 15.900 & 0.100 & -20.600 \end{pmatrix} \times 10^{-4}. \quad (42)$$

**SS.** A strand-symmetric rate matrix is one which is symmetric under simultaneous interchange of nucleotides  $A$  with  $T$  and  $C$  with  $G$ . Strand symmetry imposes certain constraints on the parameters listed in Eq. (12),

namely,

$$\begin{aligned}\pi_C &= \pi_G = 0.5 - \pi_A, \\ C_{AC} &= C_{GT}, \quad C_{AG} = C_{CT}, \\ \Phi_{AC} &= \Phi_{GA}, \quad \Phi_{CG} = 0,\end{aligned}\tag{43}$$

which reduces the number of independent parameters to six, as required of a strand-symmetric matrix [12]. Most genomic sequences, when examined on a sufficiently large scale are observed to be strand-symmetric. The parameter choices  $\pi_a$  and  $C_{ab}$  in the right-hand column of Table 1 are obtained by averaging pairs of parameters constrained to be equal by Eq. (43). The one independent flux was then chosen arbitrarily within the range allowed by the constraint Eq. (13). The resulting rate matrix is

$$Q^{\text{SS}} = \begin{pmatrix} -11.231 & 2.154 & 7.231 & 1.846 \\ 1.143 & -18.000 & 0.571 & 16.286 \\ 16.286 & 0.571 & -18.000 & 1.143 \\ 1.846 & 7.231 & 2.154 & -11.231 \end{pmatrix} \times 10^{-4}. \tag{44}$$

**SSR.** If the one remaining flux, namely  $\Phi_{AC} = \Phi_{GA}$  which corresponds to a closed path  $A \rightarrow C \rightarrow T \rightarrow G \rightarrow A$  in Fig. 1(b), is constrained to be zero, the resulting rate matrix is strand-symmetric and reversible. Such a matrix has five free parameters. Setting  $\Phi_{AC} = \Phi_{GA} = 0$  while retaining the remaining parameters in the second column of Table 1 yields the strand-symmetric, reversible matrix

$$Q^{\text{SSR}} = \begin{pmatrix} -11.231 & 1.385 & 8.000 & 1.846 \\ 2.571 & -18.000 & 0.571 & 14.857 \\ 14.857 & 0.571 & -18.000 & 2.571 \\ 1.846 & 8.000 & 1.385 & -11.231 \end{pmatrix} \times 10^{-4}. \tag{45}$$

For each of the above four rate matrices a total of 1000 synthetic datasets each sampled at  $10^5$  independent genomic sites were constructed using the procedure outlined for the  $K = 3$  case in the previous section, except that the numerical experiment was carried out assuming a sample of  $M = 8$  individuals. Maximum likelihood estimates of the parameters in the rate matrices  $Q^{\text{GRM}}$  and  $Q^{\text{GTR}}$  were calculated for each dataset using the theory of Section 4. Maximum likelihood estimates of the parameters of the rate matrices  $Q^{\text{SS}}$  and  $Q^{\text{SSR}}$  were calculated using analogous likelihood functions constrained by Eqs. (43). Histograms of the estimated parameters are plotted in Figs. 4 and 5.

In common with the  $K = 3$  case, estimates of the parameters  $\pi_i$  and  $C_{ab}$  defining the reversible part are independent of  $\Phi_{ij}$ , as expected. Also in common with the  $K = 3$  case we observe that the  $\hat{\pi}_i$  estimates are centred

Table 1: Rate matrix parameters used in  $K = 4$  numerical simulations. The conventions of Eq. (12) are used with the indices 1 to 4 representing the nucleotides  $A, C, G$  and  $T$  respectively. The values of  $\Phi_{ij}$  in the table refer to the non-reversible matrices  $Q^{\text{GRM}}$  and  $Q^{\text{SS}}$  only. For the reversible matrices  $Q^{\text{GTR}}$  and  $Q^{\text{SSR}}$  all  $\Phi_{ij}$  are zero.

	GRM and GTR	SS and SSR
$\pi_A$	0.400	0.325
$\pi_C$	0.300	0.175
$\pi_G$	0.050	0.175
$C_{AC}$	$1.5 \times 10^{-4}$	$0.9 \times 10^{-4}$
$C_{AG}$	$1.6 \times 10^{-4}$	$5.2 \times 10^{-4}$
$C_{AT}$	$1.2 \times 10^{-4}$	$1.2 \times 10^{-4}$
$C_{CG}$	$0.2 \times 10^{-4}$	$0.2 \times 10^{-4}$
$C_{CT}$	$8.8 \times 10^{-4}$	$5.2 \times 10^{-4}$
$C_{GT}$	$0.3 \times 10^{-4}$	$0.9 \times 10^{-4}$
$\Phi_{CG}$	$0.15 \times 10^{-4}$	0
$\Phi_{GA}$	$-0.10 \times 10^{-4}$	$0.50 \times 10^{-4}$
$\Phi_{AC}$	$1.00 \times 10^{-4}$	$0.50 \times 10^{-4}$

about their true value, but the estimates  $\hat{C}_{ab}$  tend to be low. A similar pattern was observed by Vogl and Bergman for  $K = 2$  simulations with selection (see Fig. 3 of [6]). Estimates  $\hat{\phi}_{ij}$  of the parameters determining the extent of non-reversibility are centred about their true values, or slightly skewed towards zero. Figure 6 shows qq-plots of 1-sided p-values calculated from likelihood ratio test statistics analogous to Eq. (40) against a uniform distribution, assuming the null hypothesis  $\Phi_{ij} = 0$ . For the otherwise unconstrained reversible and non-reversible rate matrices  $Q^{\text{GTR}}$  and  $Q^{\text{GRM}}$ , the p-values were calculated assuming a chi-squared distribution with 3 degrees of freedom for the likelihood ratio test statistic. For the strand-symmetric rate matrices  $Q^{\text{SSR}}$  and  $Q^{\text{SS}}$  a chi-squared distribution with 1 degree of freedom was assumed. For the reversible matrices the p-values have a uniform distribution as required, while the p-values for the non-reversible rate matrices are suitably small.

## 6 Discussion and Conclusions

We have demonstrated that it is possible, in principle, to estimate an evolutionary rate matrix from the site frequency spectrum of an alignment of genomes sampled from a population, provided certain conditions are met.



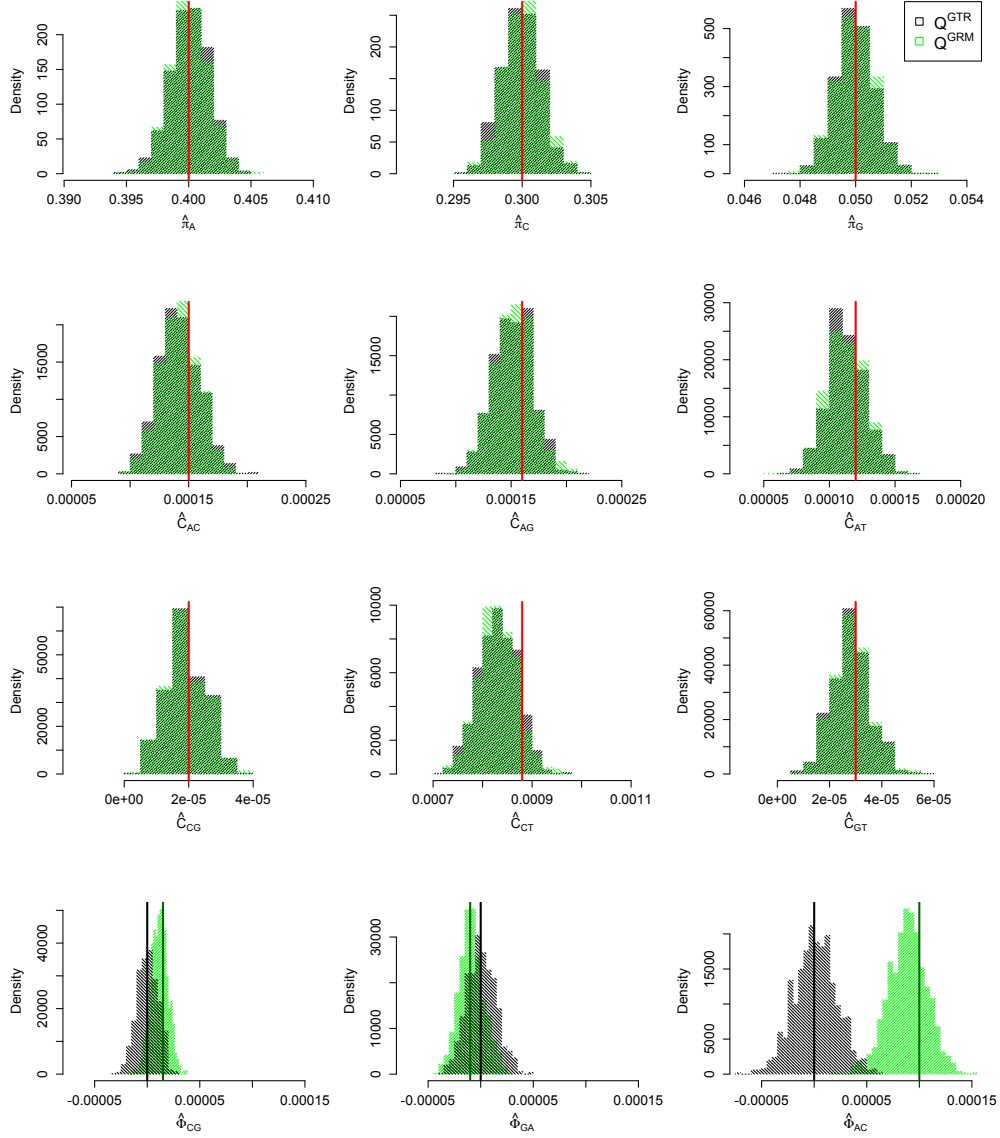


Figure 4: Histograms of estimates of the parameters  $\pi_i$ ,  $C_{ab}$  and  $\Phi_{ij}$  defining the rate matrices  $Q^{\text{GTR}}$  (black histograms) and  $Q^{\text{GRM}}$  (green histograms) for  $K = 4$  alleles. True values of the parameters are given in the first column of Table 1 and are indicated by the thick vertical lines. 1000 independent datasets were generated for each rate matrix assuming the data to be sampled from  $M = 8$  independently chosen individuals from a defining population of  $N = 30$  individuals.

The procedure consists of a maximum likelihood estimate of all 12 param-

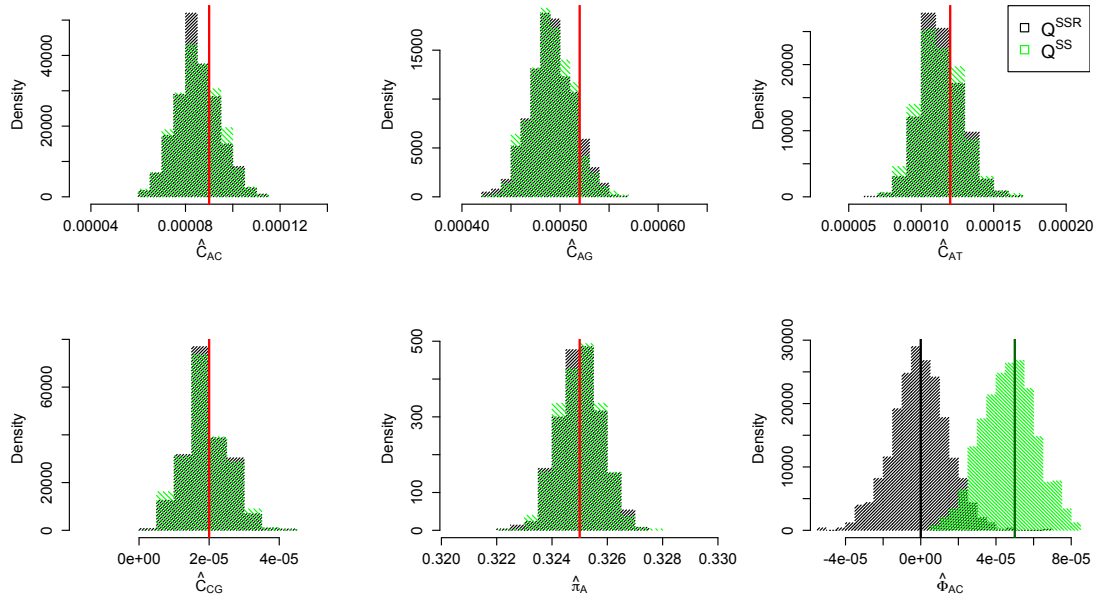


Figure 5: Histograms of estimates of the parameters  $\pi_A$ ,  $C_{AC}$ ,  $C_{AG}$ ,  $C_{AT}$ ,  $C_{CG}$  and  $\Phi_{AC}$  defining the strand-symmetric rate matrices  $Q^{\text{SSR}}$  (black histograms) and  $Q^{\text{SS}}$  (green histograms) for  $K = 4$  alleles. True values of the parameters are given in the second column of Table 1 and are indicated by the thick vertical lines. 1000 independent datasets were generated for each rate matrix assuming the data to be sampled from  $M = 8$  independently chosen individuals from a defining population of  $N = 30$  individuals.

ters of the  $4 \times 4$  nucleotide mutation rate matrix based on the distribution of observed single nucleotide polymorphisms at neutral sites in a multiple alignment. Given the site frequency spectrum constructed from the alignment of a large number of neutral sites obtained from sequencing a moderate number of individuals, the calculation is computationally straightforward and also provides a likelihood ratio test of the significance of the non-reversible part of the rate matrix.

We feel it is important to visit in turn each of the conditions required of our procedure to highlight the remaining challenges inherent in the approach.

Firstly we have assumed the population to have a constant size and to evolve according to the dynamics of a Wright-Fisher model. Except in papers specifically addressing the point, this assumption is almost universally made implicitly throughout both the population dynamics and phylogenetics literature. In practice, though, the assumption is often violated for real

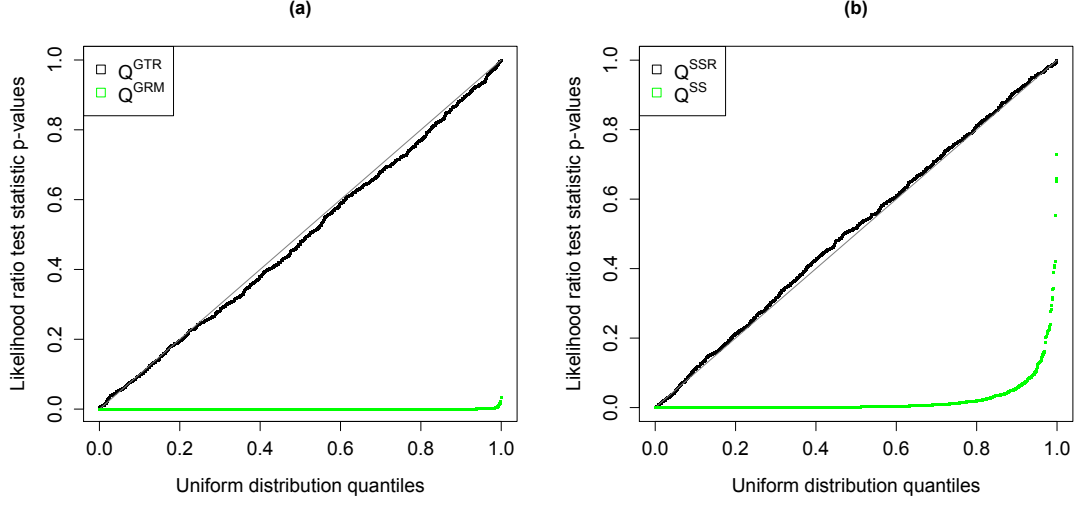


Figure 6: QQ plots of 1-sided p-values calculated from the likelihood ratio statistic assuming a null hypothesis  $\Phi_{ij} = 0$ . (a) For the matrices  $Q^{GTR}$  and  $Q^{GRM}$  likelihoods are maximised without constraints on any parameters, and p-values are calculated from the the likelihood ratio statistic assuming a chi-squared distribution with 3 degrees of freedom. (b) For the matrices  $Q^{SSR}$  and  $Q^{SS}$  likelihoods are maximised under the constraints of Eq. (43) and p-values are calculated assuming a chi-squared distribution with 1 degree of freedom. The p-values are expected to have a uniform distribution for the reversible matrices  $Q^{GTR}$  and  $Q^{SSR}$ , which satisfy the null hypothesis, but not for the non-reversible matrices  $Q^{GRM}$  and  $Q^{SS}$ .

datasets. It is important to recognise that a non-constant population size can alter the dynamics of genetic drift in ways which can effect the site frequency spectrum. For instance the recent explosive growth in the human population is believed to have skewed the site frequency spectrum towards rare numbers of derived alleles [13]. On the other hand, if a rapidly growing population is modelled with Galton-Watson branching process, alleles present in the population at the beginning of the growth period can remain endemic in the population indefinitely without dissipating through genetic drift, even in the absence of new mutations at the spectrum boundary [14]. We therefore caution that, as a general principle, the constant population Wright-Fisher model may be inappropriate for estimating mutation rates from populations of rapidly varying size.

Secondly we have assumed that a large number  $L$  of sites within a genome can be identified which have evolved independently and neutrally. Promising candidates are short intron sites [15] and 4-fold degenerate sites within codons, which are assumed to be relatively free of selective pressures. Vogl

and Bergman [6] have estimated selection effects in datasets consisting the short intron sites and 4-fold degenerate sites of a multiple alignment of 10 whole genome *Drosophila simulans* genomes. By partitioning the set of nucleotides into two effective alleles,  $\{A, T\}$  and  $\{C, G\}$ , they are able to reduce the problem of estimating the parameters of a  $K = 2$  Wright-Fisher model with both mutation and selection included, and find clear evidence of directional selection favouring the  $\{C, G\}$  state. Therefore our analysis is not suitable for this particular *D. simulans* dataset, for instance. For consistency, any analysis of a real dataset to determine the full set of 12 parameters of a  $4 \times 4$  mutation matrix would first have to pass Vogl and Bergman’s test of neutrality.

Thirdly we have assumed that the genomic sites considered evolve with a common rate matrix. There is clear evidence however that biochemical effects can render mutation rates context dependent [16, 17, 18], that is, rates may depend on the identity of neighbouring nucleotides. A strong example of this is the effect on  $C \rightarrow T$  transitions of the *CpG* context. Assuming the context of a neutral site to be subject to selection pressures and therefore relatively stable, one could possibly accommodate context dependence by restricting the dataset, and hence the estimated rate matrix, to genomic sites with a particular context.

Finally we mention the caveat that the estimation procedure relies on an approximate solution to the forward Kolmogorov equation valid in the limit of small scaled mutation rates, by which we essentially mean the Ewen-Watterson ‘ $\theta$ -parameter’ of order  $u_{ab}N$ , where  $N$  is the population size and  $u_{ab}$  the per-generation mutation rate from allele  $A_a$  to allele  $A_b$ . As a rule of thumb, the approximate solution agrees well with numerical simulations of the neutral Wright-Fisher model provided  $\theta < O(10^{-2})$  [1].

## References

- [1] C. J. Burden, Y. Tang, An approximate stationary solution for multi-allele neutral diffusion with low mutation rates (July 2016). **arXiv:** 1607.00104.  
URL <http://arxiv.org/abs/1607.00104>
- [2] G. Watterson, On the number of segregating sites in genetical models without recombination, *Theoretical population biology* 7 (2) (1975) 256–276.
- [3] W. Ewens, A note on the sampling theory for infinite alleles and infinite sites models, *Theoretical population biology* 6 (2) (1974) 143–148.
- [4] A. RoyChoudhury, J. Wakeley, Sufficiency of the number of segregating sites in the limit under finite-sites mutation, *Theoretical population biology* 78 (2) (2010) 118–122.

- [5] C. Vogl, Estimating the scaled mutation rate and mutation bias with site frequency data, *Theoretical population biology* 98 (2014) 19–27.
- [6] C. Vogl, J. Bergman, Inference of directional selection and mutation parameters assuming equilibrium, *Theoretical population biology* 106 (2015) 71–82.
- [7] C. Lanave, G. Preparata, C. Saccone, G. Serio, A new method for calculating evolutionary substitution rates, *Journal of molecular evolution* 20 (1) (1984) 86–93.
- [8] S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences, *Lectures on mathematics in the life sciences* 17 (1986) 57–86.
- [9] A. Etheridge, Some Mathematical Models from Population Genetics: École D’Été de Probabilités de Saint-Flour XXXIX-2009, Vol. 2012 of *Lecture Notes in Mathematics*, Springer, Berlin Heidelberg, 2011.
- [10] M. Cao, J. Shi, J. Wang, J. Hong, B. Cui, G. Ning, Analysis of human triallelic snps by next-generation sequencing, *Annals of human genetics* 79 (4) (2015) 275–281.
- [11] C. Phillips, J. Amigo, Á. Carracedo, M. Lareu, Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data, *Forensic Science International: Genetics* 19 (2015) 100–106.
- [12] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, *Journal of Molecular Evolution* 40 (3) (1995) 318–325.
- [13] A. Keinan, A. G. Clark, Recent explosive human population growth has resulted in an excess of rare genetic variants, *Science* 336 (6082) (2012) 740–743. [arXiv:http://www.sciencemag.org/content/336/6082/740.full.pdf](http://www.sciencemag.org/content/336/6082/740.full.pdf), doi:10.1126/science.1217283.  
URL <http://www.sciencemag.org/content/336/6082/740.abstract>
- [14] C. J. Burden, H. Simon, Genetic drift in populations governed by a galton–watson branching process, *Theoretical population biology* 109 (2016) 63–74.
- [15] J. Parsch, S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, P. Andolfatto, On the utility of short intron sequences as a reference for the detection of positive and negative selection in drosophila, *Molecular biology and evolution* 27 (6) (2010) 1226–1234.
- [16] Z. Zhao, E. Boerwinkle, Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome, *Genome research* 12 (11) (2002) 1679–1686.

- [17] A. Siepel, D. Haussler, Phylogenetic estimation of context-dependent substitution rates by maximum likelihood, *Molecular Biology and Evolution* 21 (3) (2004) 468–488.
- [18] D. G. Hwang, P. Green, Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution, *Proceedings of the National Academy of Sciences of the United States of America* 101 (39) (2004) 13994–14001.